Source/Filter Factorial Hidden Markov Model, with Application to Pitch and Formant Tracking

Jean-Louis Durrieu and Jean-Philippe Thiran, *Senior Member, IEEE*

Abstract

Tracking vocal tract formant frequencies (f_p) and estimating the fundamental frequency (f_0) are two tracking problems that have been tackled in many speech processing works, often independently, with applications to articulatory parameters estimations, speech analysis/synthesis or linguistics. Many works assume an auto-regressive (AR) model to fit the spectral envelope, hence indirectly estimating the formant tracks from the AR parameters. However, directly estimating the formant frequencies, or equivalently the poles of the AR filter, allows to further model the smoothness of the desired tracks. In this paper, we propose a Factorial Hidden Markov Model combined with a vocal source/filter model, with parameters naturally encoding the f_0 and f_p tracks. Two algorithms are proposed, with two different strategies: first, a simplification of the underlying model, with a parameter estimation based on variational methods, and second, a sparse decomposition of the signal, based on Non-negative Matrix Factorization methodology.

The results are comparable to state-of-the-art formant tracking algorithms. With the use of a complete production model, the proposed systems provide robust formant tracks which can be used in various applications. The algorithms could also be extended to deal with multiple-speaker signals.

Index Terms

Speech analysis, speech synthesis, formant tracking, source/filter model, variational methods, expectation-maximization (EM) algorithm, non-negative matrix factorization (NMF).

The authors are with the Signal Processing Laboratory (LTS5), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, e-mail: firstname.lastname@epfl.ch.

This work was funded by the Swiss CTI agency, project n. 11359.1 PFES-ES, in collaboration with SpeedLingua SA, Geneva, Switzerland.

I. Introduction

Tracking the formants (or the Vocal Tract Resonances - VTR) has been extensively investigated, with applications ranging from speech synthesis, or speech recognition, to higher level linguistics. We first review some related works, then discuss the contributions of the proposed model and algorithms. The organization of this article is provided at the end of this introduction.

A. Related works

The source/filter model has been the prevailing model since its introduction by Fant [1]: the speech signal is produced from the excitation of the vocal tract by the glottal source signal. The formants are the local maxima of the vocal tract frequency response magnitude.

The estimation of these formants, through their frequency and amplitude, allows several applications such as formant-based speech synthesis or acoustical analysis of speech signals for linguistic purposes [2]. It has also been considered for speech recognition [3]. Formant estimation is an interesting topic that permits insightful analysis of speech production. Assuming P formants, for a given speech segment, if the formant frequencies f_p correspond to the vocal tract resonances, then the vocal tract position and shape could be retrieved. Tracking these frequencies through time could further allow inference of articulatory movements. The results can then be applied for medical or pedagogic purposes, correcting deficient pronunciations or pronunciation training for foreign languages [4].

Most formant estimation systems consist of two steps: first, the filter frequency response is estimated, and second, the peaks of that frequency response are identified and tracked. The filter estimation usually relies on the estimation of the Linear Predictive Coefficients (LPC) (or, equivalently, the Auto-Regressive (AR) coefficients) as in the WaveSurfer software, with the Entropic Signal Processing System (ESPS) programs [5] or with cepstral representations [6], [7]. From the AR coefficients, it is easy to compute the formant frequency tracks: they are the frequencies of the poles of the frequency filter. However, for several cases, the estimation needs to be refined: two poles with close frequencies are likely to be merged, and spurious formants may result from low amplitude poles. Some other works [8], [9] propose to compute the filter response by approximating the spectrum with a parameterized probability density function. In those cases, both the formant frequency candidate selection and the tracking can be jointly done.

In order to address the aforementioned issues, several tracking strategies have been proposed. Most of these assume a Markov model on the frequency sequence. In the software Praat, each formant track is estimated independently [7]: for each formant, a “central” frequency value, the frequency and bandwidth

costs as well as the transition cost have to be provided, and are used in a Viterbi algorithm. The underlying model takes as observation likelihood a weighted sum of two functions, respectively depending on the frequency and the bandwidth of the formant candidate. The transition between two consecutive frames depends only on the relative difference between the logarithm of the frequency values. The f_p tracks are estimated independently one from the other. A similar model can be seen in [10], where the authors propose a partial tracking strategy that favors the continuity of the slope of the tracks, rather than the continuity in frequency. It is also reported to work well on formant tracking. A notable difference to the previous work comes from the estimation process, which pursues all the tracks at once, hence limiting duplicates in the estimation. The state model and Kalman filter approaches in [11], [12], [13] also provide such a simultaneous formant estimation.

It is also interesting to mention the method in [14], which relies on minimizing the output of inverse filtering the speech signal, assuming filters related with the AR speech model. Such an algorithm therefore directly relies on the original time-domain signal, hence avoiding any approximation related with the estimation of features such as the LPCs or the cepstral coefficients. Additionally, two analysis-by-synthesis methods have striking similarities to the algorithms proposed in this article: in [15], [16], the authors propose to estimate the articulatory and AR parameters, respectively, by decomposing the signal onto a dictionary of elements generated with different values of the desired quantities.

At last, in this article, the proposed approaches mostly rely on Factorial Hidden Markov Models (FHMM) as used by many works on source separation [17], [18], [19], as well as on Non-negative Matrix Factorization (NMF) [20], [21]. However, in these works, the signal is composed of different sources to be separated, while we focus here on one single source, the human voice, and model the many different variabilities by as many different "sources" which describe the speech production process.

B. Contributions

In order to model the smoothness in the evolution of the formant frequencies, we propose to use an FHMM in combination with a source/filter model. The purpose is to completely model the input speech signal within a unified model: it therefore includes the fundamental frequency (F0), the vocal tract filter with the corresponding formants, as well as compensations such as the energy and channel compensations. This model is referred to as a Source/Filter FHMM (SFFHMM).

The proposed SFFHMM however corresponds to an HMM with far too many states to allow the use of standard estimation algorithms, such as the forward-backward algorithm (or Baum-Welch algorithm) [22]. Instead, two approximations are studied. The first one is derived from variational approaches, and relies

on the use of an alternative likelihood, which is simpler in the sense that quantities of interest (such as sufficient statistics) can be computed with a relatively low complexity. In the log-spectral domain, we show that the problem of the source and filter parameters estimation turns out to be a linear model, therefore allowing the use of the variational method for FHMMs derived in [23]. The second approach uses more heuristics and consists in modifying the model down to a problem similar to an NMF problem, as was introduced in [21]. In this article, that algorithm is presented as an approximation to the "true" SFFHMM model, and is further extended, in particular to take into account recording conditions.

C. Organization

This article is organized as follows: the proposed exact model is first described. Thereafter, the variational approximation and algorithm are presented, followed by the NMF-based approximation. We then discuss the results of the proposed algorithms on publicly available datasets, and conclude on perspectives for this work.

II. Source/Filter Factorial Hidden Markov Model

In this section, the statistical signal model is described: the quantities of interest, the pitch F0 and the formants, are considered as hidden variables, each of which is further modeled as a Markov chain. The likelihood of the observed signal given these latent variables is first described in Sect. II-A. The evolution models for the formants and F0 are then discussed in Sect. II-B. At last, the practical choices that center the model back onto the context of speech processing, are presented in Sect. II-C.

A. Source/Filter model: hidden finite state formulation

The observed signal is a clean speech segment, with a single speaker. Assuming the source/filter model from [1], the time-domain speech signal x is produced as a glottal source signal filtered by the vocal tract filter [21][1]. Let s_{fn} be the power spectrum density (PSD) at frequency bin f such that $s_{fn} \approx |x_{fn}|^2$, with $\mathbf{x}_n = [x_{1n}, \ldots, x_{Fn}]^T$ the Fourier transform of size F of a frame $n = \in \{1, \ldots, N\}$. The number N denotes the number of analysis frames of x. The PSD s_{fn} is approximated by a parameterized version $\bar{s}_{fn}$ as given by the product, in the frequency domain, between the filter contribution s^{Φ}_{fn} (the squared absolute frequency response of the underlying vocal tract filter) and the source PSD $s^{F_0}_{fn}$:

$$s_{fn} \approx \bar{s}_{fn} = s^{\Phi}_{fn} s^{F_0}_{fn} \tag{1}$$

[1]In this paper, the following notation convention is adopted: scalar values are generally denoted with lower case letters (*e.g.* s or s_{fn}), vectors by bold lower case letters ($\mathbf{s}_n$) and tensors of 2 or more dimensions by bold upper case letters ($\mathbf{S}$).

When the source or excitation signal is voiced, it is mainly parameterized by the fundamental frequency f_0, hence the super-script F_0.

As in [12], the filter part s_{fn}^{Φ} is assumed to follow an all-pole model. Let z_{pn}, for $p = 1, \ldots, P$, be its distinct poles, *i.e.* the real and complex conjugate pairs of poles, where P is the number of formants. For a pole $z_{pn} = \rho_{pn} \exp \frac{2j\pi f_{pn}}{F_s}$, where F_s is the sampling rate of x, f_{pn} is referred to as the p^{th} formant frequency, at frame n, and ρ_{pn} its amplitude. For complex-only poles z_{pn}, we define $s_{fn}^{F_p} = \left|(1 - z_{pn}e^{-2j\pi f/F})(1 - z_{pn}^{*}e^{-2j\pi f/F})\right|^{-2}$. The filter power spectrum, for formant F_p, $\mathbf{s}_n^{F_p} = [s_{1n}^{F_p}, \ldots, s_{Fn}^{F_p}]^T$ is therefore the squared magnitude of the frequency response of the order-2 AR (AR(2)) filter, with poles $\{z_{pn}, z_{pn}^{*}\}$. Thereafter, $s_{fn}^{\Phi} = \prod_{p=1}^{P} s_{fn}^{F_p}$ and:

$$s_{fn} \approx \bar{s}_{fn} = \left(\prod_{p=1}^{P} s_{fn}^{F_p} \right) s_{fn}^{F_0} \tag{2}$$

The proposed framework enables to jointly estimate the spectral envelope $\prod_{p=1}^{P} s_{fn}^{F_p}$, the pitch f_{0n} and the formant frequencies f_{pn}. It aims at allowing formant frequency tracking directly during the envelope estimation process. With mild smoothness assumptions, such a framework allows to deal with merged f_p values.

In order to do so, instead of estimating the f_{pn} values, for each p and n, or estimating the LPC coefficients, we generate several AR(2) filter frequency responses (for $p = 1 \ldots P$) with different fixed values of (f_p, ρ_p) and several glottal source power spectra (for $p = 0$) with different f_0 values. Each $\mathbf{s}_n^{F_p}$ is then equal to one of these elements in the corresponding power spectrum dictionary, an approach very similar to the sparse decomposition proposed in [16].

In this work, the variables of interest are the formant tracks, *i.e.* the sequences of frequencies $\{f_{p1}, \ldots, f_{pN}\}, \forall p \in [0, P]$. For $p = 0 \ldots P$, let K be the number of elements in the dictionary $\mathbf{W}^p = [\mathbf{w}_1^p, \ldots, \mathbf{w}_K^p]$ (further detailed in Sect. II-C). Let $\mathbf{g}_n^p = [g_{1n}^p, \ldots, g_{Kn}^p]^T$ be the state vector, such that $g_{kn}^p = 0$, $\forall k$, except for the active state at frame n, for which $g_{kn}^p = 1$. With these quantities, we have:

$$\begin{aligned} \mathbf{s}_n^{F_p} &= \mathbf{W}^p \mathbf{g}_n^p \\ \mathbf{S}^{F_p} &= \mathbf{W}^p \mathbf{G}^p \end{aligned} \tag{3}$$

$$\mathbf{S} \approx (\mathbf{W}^1 \mathbf{G}^1) \bullet \ldots \bullet (\mathbf{W}^P \mathbf{G}^P) \bullet (\mathbf{W}^0 \mathbf{G}^0) \tag{4}$$

where $\bullet$ is the Hadamard product, and $\mathbf{S}$ the $F \times N$ PSD matrix of general element s_{fn}. Eq. (4) results from replacing the formant and F0 contributions $s_{fn}^{F_p}$ in Eq. (2) by their corresponding expressions in (3).

Since the spectra in the $\mathbf{W}^p$ are fixed, we need to compensate several effects: the energy of the observed signal, first, by multiplying (2) by an energy parameter h_n. Second, the recording conditions need to be compensated for, by means of an additional filter frequency response r_f. Let $\mathbf{r} = [r_1, \ldots, r_F]^T$, the model therefore becomes:

$$\mathbf{s}_n \approx h_n \mathbf{r} \bullet (\mathbf{W}^1 \mathbf{g}_n^1) \bullet \ldots \bullet (\mathbf{W}^P \mathbf{g}_n^P) \bullet (\mathbf{W}^0 \mathbf{g}_n^0) \tag{5}$$

We now assume that the Short-Term Fourier Transform (STFT) of the observation, $\mathbf{X} = \{x_{fn}\}_{fn}$, conditionally upon the state vectors $\{\mathbf{g}_n^p\}_p$, is distributed as follows [20]: each frame $\mathbf{x}_n$ is a complex proper Gaussian vector, centered, with a diagonal covariance[2]. If the underlying time-domain signal x is wide-sense stationary (w.s.s.), then the diagonal of this covariance can be considered as the power spectral density (PSD) of the signal, which is equal to $\bar{\mathbf{s}}_n$ as defined in (2). Using (5), the observation likelihood given the state vectors concatenated as one vector $\mathbf{g}_n$, is distributed as:

$$\mathbf{x}_n | \mathbf{g}_n \sim \mathcal{N}_c(\mathbf{0}, \mathrm{diag}\left\{ h_n \mathbf{r} \bullet \prod\nolimits_p \mathbf{W}^p \mathbf{g}_n^p \right\}) \tag{6}$$

where $\prod_p$ is here meant as the product of Hadamard of the operands, over variable p. From the definition of the complex proper Gaussian distribution (*e.g.* in [20]), Eq. (6) explicitly becomes, with $s_{fn} = |x_{fn}|^2$:

$$\begin{aligned} \log p(\mathbf{x}_n | \mathbf{g}_n) = & -F \log \pi - \sum_f \log h_n r_f \\ & - \sum_{f,p} \log \sum_k w_{fk}^p g_{kn}^p \\ & - \sum_f \frac{s_{fn}}{h_n r_f \prod_p \sum_k w_{fk}^p g_{kn}^p} \end{aligned} \tag{7}$$

As in [20], it is interesting to draw the equivalence between maximizing (7) and minimizing the Itakura-Saito (IS) divergence between the observed variance $\mathbf{S}$ and the parameterized PSD $\bar{\mathbf{S}} = \mathbf{H} \bullet \mathbf{R} \bullet \prod_p \mathbf{W}^p \mathbf{G}^p$ in (5), with respect to the model parameters:

$$\begin{aligned} D_{\mathrm{IS}}(\mathbf{S} | \bar{\mathbf{S}}) = \sum_{f,n} & - \log \frac{s_{fn}}{h_n r_f \prod_p \sum_k w_{fk}^p g_{kn}^p} \\ & + \frac{s_{fn}}{h_n r_f \prod_p \sum_k w_{fk}^p g_{kn}^p} - 1 \end{aligned} \tag{8}$$

where $\mathbf{H}^T = \mathbf{h}[\underbrace{1, \ldots, 1}_{F}]$ and $\mathbf{R} = \mathbf{r}[\underbrace{1, \ldots, 1}_{N}]$. This equivalence explicates how to interpret the approximation sign from Eq. (5): the parameters have to be estimated such that the model PSD $\bar{\mathbf{S}}$ is as close to

[2]A diagonal covariance in that case means that we are neglecting the coupling between the frequency bins of the Fourier transforms.

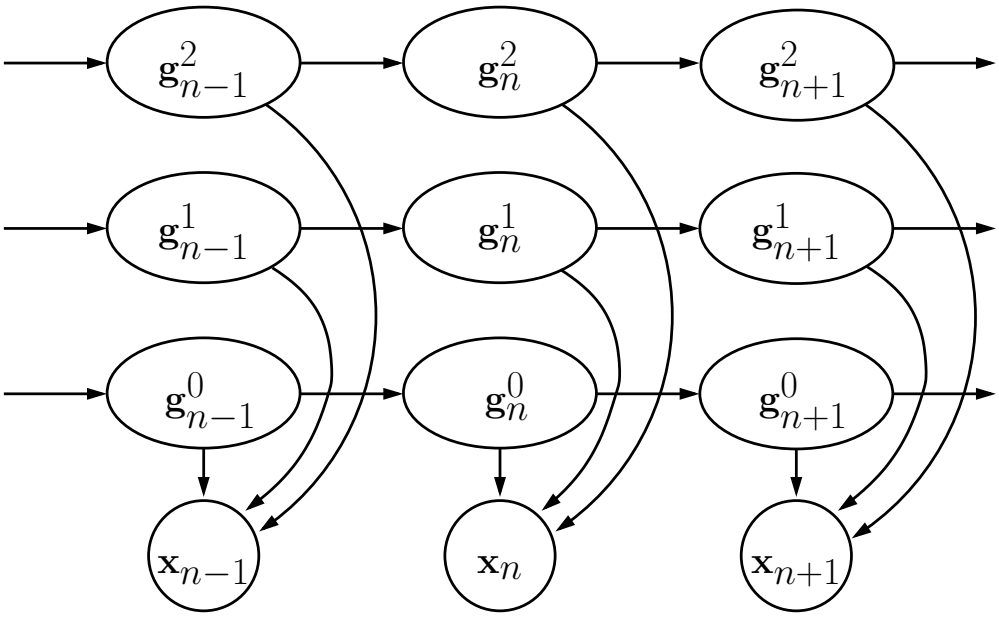


Fig. 1. Graphical model of an FHMM.

the observed power spectrum $\mathbf{S}$ as possible, in the sense of minimizing the divergence (8). In particular, the scale invariance of the IS divergence [20] gives equal importance to low and to high energy time-frequency components. The advantage is, for instance, to process the whole frequency range equally, without the need for pre-processing steps such as pre-amplification.

In order to closely fit speech signals, the dictionaries $\mathbf{W}^p$ must however overlap, in particular because two formant frequency ranges may overlap, as discussed in Sect. II-C. The likelihood (7) may therefore have several local maxima, and additional constraints on the temporal evolution of the formants is necessary, as introduced in the following section.

B. Individual Markov chains for each f_p

Owing to the observation that a human can utter different phonemes, with various combinations of formant frequencies and at different pitches, it is assumed that the sequences for the F0, the formant frequencies and the formant amplitudes can evolve independently one from another: each F0 and formant state sequence can therefore naturally be modeled as a Markov chain independent from the other chains, as represented by the graphical model in Fig. 1. There are therefore $P+1$ independent Markov chains,

each with K possible states per frame, such that:

$$p(\mathbf{G}) = \prod_{p=0}^{P} p(\mathbf{g}_1^p) \prod_{n=2}^{N} p(\mathbf{g}_n^p | \mathbf{g}_{n-1}^p) \tag{9}$$

The state sequence probability is therefore the product of the probabilities for each Markov chain, hence the name of the model, *factorial HMM* [23]. It is equivalent to an HMM, with K^P states, except that the transition probabilities are more constrained in the case of the FHMM. One should note that exact inference with K^P HMM states is prohibitively expensive, hence the approximations proposed in [23] and those developed in Sect. III and IV.

The joint probability for the observation and the states is therefore given by:

$$p(\mathbf{X}, \mathbf{G}) = \prod_{n=1}^{N} p(\mathbf{x}_n | \mathbf{g}_n) \prod_{p=0}^{P} p(\mathbf{g}_1^p) \prod_{n=2}^{N} p(\mathbf{g}_n^p | \mathbf{g}_{n-1}^p) \tag{10}$$

We further assume that the transition probabilities only depend on the difference between the logarithm of the formant or glottal source frequencies f_p. These probabilities can therefore be learned easily from an annotated database, as discussed in Sect. V-B.

At last, the desired F0 and formant tracks are given by the following maximum a posteriori (MAP) sequence $\widehat{\mathbf{F}}$:

$$\widehat{\mathbf{F}} \triangleq \{f_{kp}\}_{k=\hat{k}_1^p,\ldots,\hat{k}_N^p, p=0,\ldots,P} \tag{11}$$

$$\widehat{\mathbf{K}} = \arg\max_{\mathbf{K}} p(\mathbf{K}|\mathbf{X}) \tag{12}$$

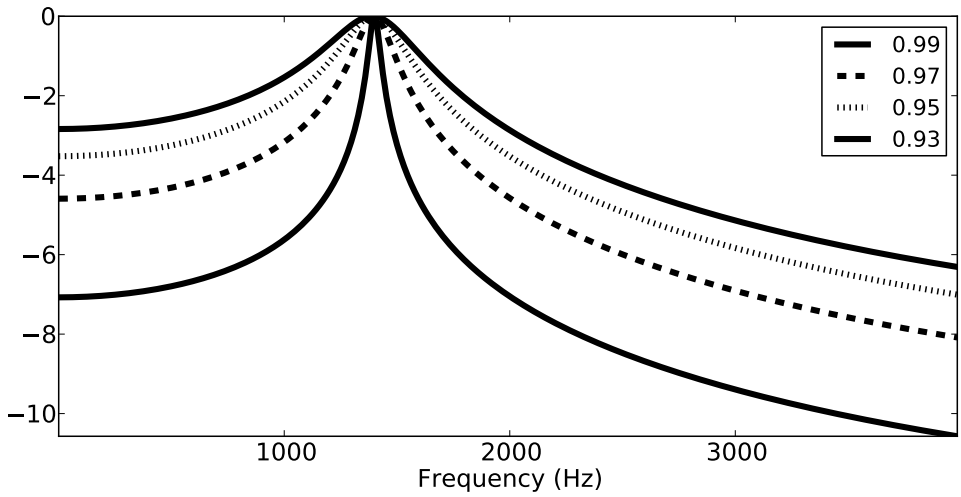
$$= \arg\max_{\mathbf{K}} p(\mathbf{G}[\mathbf{K}]|\mathbf{X}) \tag{13}$$

where $\mathbf{K} = \{k_n^p\}_{np}$ is the matrix containing the sequences of active states for each of the P chains, for all the frames $n = 1, \ldots, N$, and $\mathbf{G}[\mathbf{K}]$ is the active state tensor corresponding to $\mathbf{K}$, that is to say the tensor $\mathbf{G}$ such that $\forall (k, n, p), g_{kn}^p = 1$ if $k = k_n^p$ and 0 otherwise. The Viterbi algorithm allows to solve problem (13) without enumerating all the K^{PN} possible chains $\mathbf{K}$. This exact inference however still requires a prohibitive computational time, and the methods of Sect. III and IV are meant to provide an approximation to $\widehat{\mathbf{F}}$ at a lower computational cost.

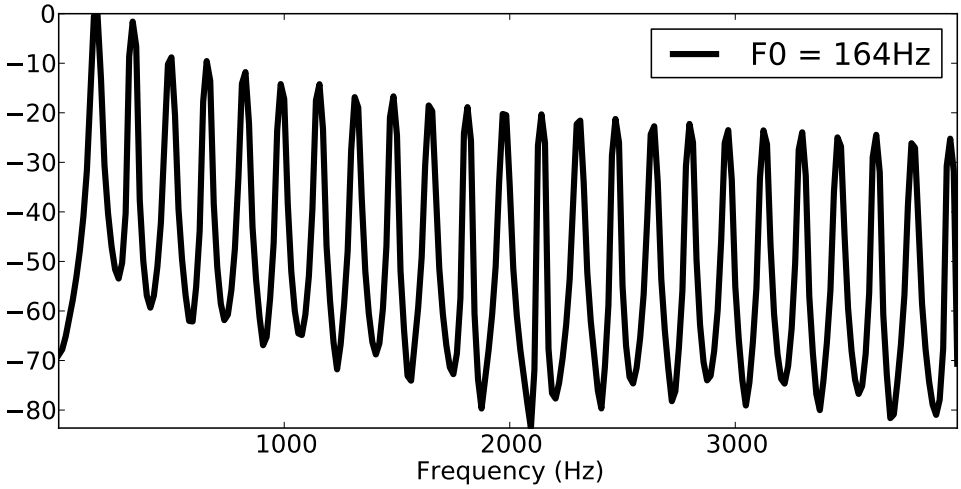
C. Choice of source/formant bases

As in [21], *the source part dictionary* $\mathbf{W}^0 = [\mathbf{w}_1^0, \ldots, \mathbf{w}_K^0]$ is fixed such that each column $\mathbf{w}_u^0$ is the power spectrum of a glottal signal at a given F0, given by $\mathcal{F}(u)$, function of the index u in the dictionary (column number):

$$\mathcal{F}(u) = F_0^{\min} \times 2^{\frac{u-1}{12U_{\text{st}}}}, \tag{14}$$



(a) Elements $\mathbf{w}_k^p$ from $\mathbf{W}^p$, same f_{kp}, different ρ_{kp}.



(b) An element $\mathbf{w}_k^0$ from $\mathbf{W}^0$.

Fig. 2. Elements from the dictionaries $\mathbf{W}^p$ (ordinates in dB).

where F_0^{min} is the lowest F0 candidate and U_{st} the number of F0 per semi-tones in the dictionary. The glottal source model KLGLOTT88 [24] was used, with the open quotient, the proportion of open glottis time during one period, arbitrarily fixed to 0.6. Ordering the columns of $\mathbf{W}^0$ by increasing F0 frequencies also allows for an easy smoothing (or tracking) procedure (as well as a convenient visual representation). Concretely, the F0 range is set to $[80, 500]$, as it corresponds to typical values for human voices. A number $U_{\text{st}} = 16$ provides a flexible dictionary, leading to about $K = 510$.

p	$F_p^{\min}$	$F_p^{\max}$	comments
0	80	500	$U_{\text{st}} = 16$
1	200	1500	adapted from [25], [2]
2	550	3500	-
3	1400	4500	-
4	2400	6000	additional values
5	3300	8000	-
6	4500	8000	-
7	5500	8000	-

TABLE I

F0 AND FORMANT FREQUENCY RANGES.

For the *filter part*, *i.e.* $p \in [1, P]$, each column $\mathbf{w}_k^p$, $p \in \{1, \ldots, K\}$, of $\mathbf{W}^p$ is the frequency response of an AR(2) filter, with complex poles $\{z_{kp}, z_{kp}^*\}$. The formant frequency range for each p is set such that $f_{kp} \in [F_p^{\min}, F_p^{\max}]$. Similarly to [16], a grid of linearly spaced values for f_{kp} and $\rho_{kp} \in [0.90, 0.99]$ is used, with G_F values for f_{kp} and G_R values for ρ_{kp}. For $g_F \in [0, G_F - 1]$ and $g_R \in [0, G_R - 1]$, we define $k = G_R g_F + g_R$ and, with $\lfloor y \rfloor$ the integer part of y:

$$f_{kp} = F_p^{\min} + (F_p^{\max} - F_p^{\min}) \lfloor k/G_R \rfloor / (G_F - 1)$$

With such a choice, elements in $\mathbf{W}^p$ that share the same frequency f_{kp} have consecutive indices k. This ordering strategy in $\mathbf{W}^p$ is important for the subsequent sparsity and smoothness-inducing procedures of Sect. IV. Elements from a dictionary $\mathbf{W}^p$ are shown on Fig. 2. Table I summarizes the chosen formant frequency and F0 ranges, inspired by [25], [2]. The sampling rate of the speech signals is 16kHz, hence the 8kHz limit for $F_p^{\max}$.

At last, in order to take into account unvoicing in the source part and the possibility of a reduction in the number of formants, the last element of each dictionary, $\mathbf{w}_K^p$ is in practice set to a "flat spectrum", such that $w_{fK}^p = 1$. For the source part, this could be interpreted as a white noise source, while for the filter part, it can be seen as a "neutral" element, which does not alter the filter response when it is selected.

Note that K is not necessarily the same for all the dictionaries. As mentioned earlier, the source dictionary $\mathbf{W}^0$ has a different number of components, depending on the F0 range as well as the value of U_{st}. Without loss of generality, and for simplicity, in the remainder of this article we assume a unique K for all the dictionaries. The different values used for G_F and G_R are given in Sect. V-C.

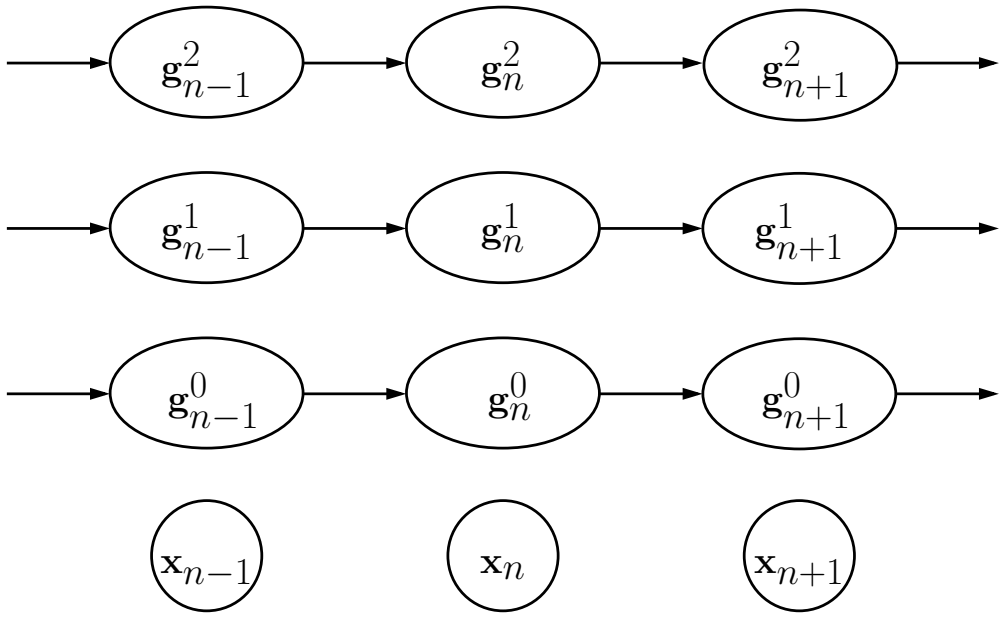


Fig. 3. Graphical model of the structured variational approximation to the FHMM 1.

III. Variational Approximation

A. Alternative Joint Probability

In order to infer the most likely sequence of states for each of the Markov chains, and because of the great complexity coming from the use of several chains, with many possible states each, approximate inference algorithms are required.

The variational approximation (VA) proposed in [23] consists in proposing of a joint probability alternative to the "true" probability (10). With that probability proposal, it is easier to compute sufficient statistics or quantities like the posterior probabilities of the FHMM states. The VA probability also depends on some additional parameters, the variational parameters, which have to be estimated in order to make the proposal closer to the actual probability, for instance by minimizing the Kullback-Leibler (KL) divergence between the VA and the true probabilities.

In [23], the structured variational approximation consists in considering the alternative model where the Markov chains are uncoupled. From a graphical model point of view, this corresponds to removing arrows going for each chain to the observation layer, as depicted on Fig. 3. The proposed probability is then:

$$q(\mathbf{G}|\theta) = \frac{1}{Z_q} \prod_{p=0}^{P} q(\mathbf{g}_1^p|\theta) \prod_{n=2}^{N} q(\mathbf{g}_n^p|\mathbf{g}_{n-1}^p, \theta) \tag{15}$$

where θ is the set of variational parameters. Z_q is the normalizing constant. Since the state indicators g_{kn}^p are either 0 or 1, the probabilities in (15) can be written as:

$$q(\mathbf{g}_1^p|\theta) = \prod_{k=1}^{K} \left(t_{k1}^p \pi_k^p\right)^{g_{k1}^p} = \sum_{k=1}^{K} g_{k1}^p t_{k1}^p \pi_k^p \tag{16}$$

$$q(\mathbf{g}_n^p|\mathbf{g}_{n-1}^p, \theta) = \prod_{k=1}^{K} \left(t_{kn}^p \sum_{j=1}^{K} P_{kj}^p g_{j(n-1)}^p \right)^{g_{kn}^p} \tag{17}$$

The variational parameters are therefore $\theta = \{\mathbf{T}^p, \pi^p, \mathbf{P}^p\}_{p=0,\dots,P}$, π^p and $\mathbf{P}^p$ are the a priori probabilities and transition probabilities for chain p. Comparing the above equations with (10), one can see that the only difference here is that the observation probabilities $p(\mathbf{x}_n|\mathbf{g}_n)$ have been replaced by the product $\prod_p(\sum_k t_{kn}^p g_{kn}^p)$: the structured VA can therefore be interpreted as factorizing the observation likelihood into likelihoods corresponding to each chain. $\mathbf{T}^p$ thereafter acts as an observation probability for chain p, and inference about the most likely state sequence or the posterior probabilities for $\mathbf{G}$ can be carried within that chain as if it were an individual HMM, with K states [23].

In order to do so, $\mathbf{T}^p$ needs to be estimated. In [23], the authors propose the KL divergence between (15) and (10) as target to be minimized:

$$D_{\mathrm{KL}}(q|p) = \sum_{\mathbf{G}} q(\mathbf{G}) \log \frac{q(\mathbf{G})}{p(\mathbf{G}|\mathbf{X})} \tag{18}$$

However, in our signal model, the derivation of the algorithm does not allow to uncouple the chains through this "trick". Indeed, due to the non-linear form of the likelihood in (7), obtaining estimates of $\mathbf{T}^p$ requires as many computations as if estimating the exact inference with the equivalent full HMM, hence a complexity in $\mathcal{O}(NK^{2(P+1)})$. Exact inference is also possible with a junction tree algorithm, but would also result in an intractable inference of $\mathcal{O}(N(P+1)K^{P+2})$ complexity [23].

It may be possible to factorize more the problem, and reduce the complexity of the estimation. However, keeping the general structure of the original model is an asset, and we investigated a linearization of our problem, in order to be able to use the derivations from [23].

B. Log Power Formulation

As stated above, using [23] for approximating the joint probability (10), with the likelihood (7) leads to an algorithm with the same time complexity as the corresponding full HMM. Another observation likelihood is therefore required. In [26], Benaroya proposed to model the signals thanks to the logarithm of their power spectra, but with the same underlying proper Gaussian model. In particular, the variance of the log-spectrum is a constant, and its mean depends on the variance parameter of the Gaussian model:

Property 1 *Let $x \sim \mathcal{N}_c(0, \sigma^2)$, and $y = \sigma^{-1}x \sim \mathcal{N}_c(0, 1)$. Then the mean and the variance of $\log|x|^2$ are given by:*

$$E[\log|x|^2] = \log\sigma^2 + E[\log|y|^2] \tag{19}$$

$$\mathit{Var}(\log|x|^2) = \mathit{Var}(\log|y|^2) \tag{20}$$

The proof is straight-forward, using the definition of the complex proper Gaussian as in Eq. (7). Numerically, we found that $E[\log|y|^2] \approx -0.57$ and $\mathrm{Var}(\log|y|^2) \approx 1.6$ (see also [27]). The above result can be extended to multivariate random variables, and we obtain, for the proposed model:

$$x_{fn}|\mathbf{g}_n \sim \mathcal{N}_c(0, h_n r_f \prod_p \sum_k w_{fk}^p g_{kn}^p) \tag{21}$$

$$\Rightarrow \begin{cases} E[\log \mathbf{s}_n|\mathbf{g}_n] & \approx & \log(h_n\mathbf{r} \bullet \prod_p \mathbf{W}^p\mathbf{g}_n^p) - 0.57 \\ Cov(\log \mathbf{s}_n|\mathbf{g}_n) & \approx & 1.6\mathbf{I} \end{cases}$$

where we recall that the observation power spectrum vector, $\mathbf{s}_n$, is defined as $s_{fn} = |x_{fn}|^2$.

The probability density function (pdf) of $\log \mathbf{s}_n|\mathbf{g}_n$ is however still difficult to use and a further approximation is necessary: the "true" pdf could be approximated by a Gaussian distribution[3], centered on $E[\log \mathbf{s}_n|\mathbf{g}_n]$ and with diagonal covariance matrix $\mathbf{C} = 1.6\mathbf{I}$. The observation model therefore becomes:

$$\log \mathbf{s}_n|\mathbf{g}_n \sim \mathcal{N}(\log(h_n\mathbf{r} \bullet \prod\nolimits_p \mathbf{W}^p\mathbf{g}_n^p) - 0.57, \mathbf{C})$$

$$\tilde{\mathbf{s}}_n|\mathbf{g}_n \sim \mathcal{N}(\tilde{h}_n + \tilde{\mathbf{r}} + \sum_p \widetilde{\mathbf{W}}^p\mathbf{g}_n^p, \mathbf{C}) \tag{22}$$

where $\tilde{\mathbf{s}}_n$, $\tilde{h}_n$, $\tilde{\mathbf{r}}$ and $\widetilde{\mathbf{W}}^p$ are the $\log$ versions of the corresponding variables. Note that the offset -0.57 has been integrated into $\tilde{h}_n$, without loss of generality. Using the model (22), it is now possible to use the VA algorithm proposed in [23].

In practice, the use of the logarithm of the power spectra tends to give too much weight to low energy components, since the logarithm tends to infinity as the energy tends to 0. The spectra $\mathbf{w}_k^0$ are generated with a sinusoidal model, which happens to create artifacts due to aliasing and windowing effects: these artifacts are minor as they are not audible, however the logarithm of $\mathbf{w}_k^0$ often shows patterns in low energy coefficients, *i.e.* those that do not correspond to harmonics of the desired F0. Preliminary experiments showed that this affected the results. A work-around for this issue is to add some noise, or equivalently an unvoiced component, b to each column component of $\mathbf{W}^0$, in particular, $\widetilde{\mathbf{w}}_k^0 \leftarrow \log(\mathbf{w}_k^0 + b)$. For these

[3]A Gaussian mixture could also be used, but this would require yet another layer of approximation, as the KL divergence (18) would still be non-linear in $\mathbf{G}$.

experiments, setting b to an arbitrary small fraction (*e.g.* $1/10000$) of the maximum value in $\mathbf{W}^0$ was found to lead to good results. It would also be possible to estimate that noise threshold b like the other energy parameters, within the EM algorithm, but we found that it led to poorer results.

C. VA Algorithm

The VA algorithm proposed in [23] is an Expectation-Maximization (EM) algorithm. It allows to compute, in the E-step, the sufficient statistics, with respect to the VA distribution q, *i.e.* the posterior probabilities $\langle \mathbf{G} \rangle \triangleq E_q[\mathbf{G}|\theta]$ and the variational parameters θ, and in the M-step, the updates of the different parameters of our model. Once the parameters for a given signal are all estimated, the optimal sequence in the sense of (13) can be computed.

The E-step consists in a loop alternating the estimation of the variational parameters θ (by minimizing (18)) with the estimation of the sufficient statistics, since in the VA, the former depends on the latter and reciprocally. The derivations and details are given in App. B. Minimizing (18) with respect to the variational parameters $\mathbf{T}^p$, we obtain:

$$\begin{aligned}\tilde{\mathbf{t}}_n^p \leftarrow (\widetilde{\mathbf{W}}^p)^T \mathbf{C}^{-1} (\tilde{\mathbf{s}}_n - \tilde{h}_n - \tilde{\mathbf{r}} \\ - \sum_{m \neq p} \widetilde{\mathbf{W}}^m \langle \mathbf{g}_n^m \rangle) - \frac{1}{2} \Delta^p \end{aligned} \tag{23}$$

where $\Delta^p = \text{diag}\left\{ (\widetilde{\mathbf{W}}^p)^T \mathbf{C}^{-1} \widetilde{\mathbf{W}}^p \right\}$. This formula is consistent with the intuition, discussed in Sect. III-A, that t_{kn}^p is like an observation probability, for chain p, given that $g_{ln}^p = 1$ for $l = k$, and 0 otherwise. Indeed, in (23), we see that $\widetilde{t}_{kn}^p$ is updated as a correlation between the spectral shape $\widetilde{\mathbf{w}}_k^p$ and the observation $\widetilde{s}_n$ from which the expected contributions of all the other chains $q \neq p$ have been removed.

In addition, since Eqs (15), (16), and (17) are formally equivalent to a single HMM chain with $\mathbf{T}^p$ as the observation likelihood, $\langle \mathbf{G}^p \rangle$ can be computed by the forward-backward algorithm as done in [23].

The M-step consists in updating the different parameters of the model, i.e. $\mathbf{h}$, $\mathbf{r}$ and, if needed, the HMM parameters π^p and $\mathbf{P}^p$. The EM criterion to be maximized in the M-step is given by:

$$\mathcal{Q}(\theta|\theta^{\text{old}}) = E_q[\log p(\mathbf{S}, \mathbf{G}|\theta)|\mathbf{S}, \theta^{\text{old}}] \tag{24}$$

where ϕ^{old} is the current set of estimated parameters. It is further detailed in App. A. Minimizing (24)

with respect to $\mathbf{h}$ and $\mathbf{r}$ (alternatively), we obtain the following updating formulas, with $\mathbf{1}_F = [\underbrace{1, \ldots, 1}_{F}]^T$:

$$\tilde{\mathbf{h}} \leftarrow \frac{1}{\mathbf{1}_F^T \mathbf{C}^{-1} \mathbf{1}_F} \mathbf{1}_F^T \mathbf{C}^{-1} (\widetilde{\mathbf{S}} - \widetilde{\mathbf{R}} - \sum_p \widetilde{\mathbf{W}}^p \langle \mathbf{G}^p \rangle) \tag{25}$$

$$\tilde{\mathbf{r}} \leftarrow \frac{1}{N} \sum_{n=1}^{N} (\tilde{\mathbf{s}}_n - \tilde{h}_n - \sum_p \widetilde{\mathbf{W}}^p \langle \mathbf{g}_n^p \rangle) \tag{26}$$

In practice, several iterations of the EM steps are necessary to converge towards stable estimates. Since the dictionaries are fixed, the number of parameters to be estimated is relatively low. The proposed EM algorithm is however sensitive to initialization. In particular, for the posterior probabilities $\langle \mathbf{G} \rangle$, several initialization schemes were tried: uniform, random or "triangle" initializations, as well as LPC initialized probabilities. With the "triangle" initialization, the VA-SFFHMM algorithm starts with probabilities that are maximum in the middle of each formant range, then linearly decreasing down to 0 at the boundaries of that range. The "LPC" method consists in setting the maximum of these probabilities around the LPC formant estimates. Preliminary results showed that the "triangle" initialization was the most promising, and this is the one which was used for the remainder of this article.

Furthermore, since $\mathbf{h}$ and $\mathbf{r}$ do not depend on the chain number p, they are also updated after the E-step for each chain. Note also that since $\mathbf{r}$ is the frequency response of a recording condition, it should be smooth in the frequency domain. This smoothness is imposed at each iteration after the update (26) by low-pass-filtering the result. It could also be constrained directly in the model, for instance by decomposing $\mathbf{r}$ on a dictionary of smooth functions, as is done for the other approach in Sect. IV, but this requires further investigations, since it leads to updating rules that are more complicated than Eq. (26).

At last, the MAP sequences of (13) are here computed as follows:

$$\widehat{\mathbf{K}} = \left\{ \widehat{\mathbf{k}}^p \right\}_p = \left\{ \arg\max_{\mathbf{k}^p} q(\mathbf{k}^p | \theta) \right\}_p \tag{27}$$

Problem (27) can be efficiently solved by a Viterbi algorithm on each of the P chains. In practice, it boils down to considering each individual HMM chain as an independent HMM, estimating the optimal sequence using the estimated variational parameters $\mathbf{T}^p$ as the conditional likelihoods, $p(\mathbf{x}_n | \mathbf{g}_n^p)$ [23].

The EM algorithm for VA of the Source/Filter FHMM (VA-SFFHMM) is given in Alg. 1. As can be seen, the complexity is given by the complexity of computing the variational parameters $\mathbf{T}^p$ and by $(P+1)$ times a forward-backward algorithm on a K-state HMM. The former complexity is NP times that of Eq. (23), which is roughly in $\mathcal{O}(KF(K+F))$. The complexity of one forward-backward iteration is in $\mathcal{O}(NK^2)$. The VA-SFFHMM computation time, for each EM iteration, seems however dominated by this latter complexity, of the order of $\mathcal{O}(N(P+1)K^2)$, because it requires operations such

as comparisons while computing $\mathbf{T}^p$ requires matrix products. At last, the complexity of problem (27) is in $\mathcal{O}(PNK)$, to be compared with that of problem (13) which is in $\mathcal{O}(NK^P)$.

Algorithm 1 VA-SFFHMM: EM algorithm for variational approximation of an SFFHMM.

Initialize the posterior probabilities $\langle \mathbf{G} \rangle$, $\mathbf{h}$ and $\mathbf{r}$.
while Estimation not converged **do**
  E-step
  while posterior probabilities not converged **do**
    for $p = 0$ to P **do**
      Update variational parameters $\mathbf{T}^p$ with (23)
      Compute $\langle \mathbf{G}^p \rangle$, via forward-backward algorithm [22], with $\mathbf{T}^p$ as observation probability
      Update $\mathbf{h}$ and $\mathbf{r}$ with Eqs. (25) and (26)
    end for
    Check convergence of $\langle \mathbf{G} \rangle$
  end while
  M-step
  Update $\mathbf{h}$ and $\mathbf{r}$ with Eqs. (25) and (26)
end while
for $p = 0$ to P **do**
  Viterbi algorithm on individual p^{th} HMM to estimate $\widehat{\mathbf{k}}^p$
end for

IV. Sparse Non-Negative Matrix Factorization

A. Principle and unconstrained NMF

Another approach, which was developed in [21], consists in relaxing the constraint on $\mathbf{G}$ by assuming that $\forall n, k, p$, $g_{kn}^p \geq 0$, with $\sum_k g_{kn}^p = 1$ (to avoid indeterminacies). The parameters are first estimated, and then refined to fulfill the required original assumptions on the signal.

Maximizing the likelihood (6), w.r.t. the parameters $\mathbf{G}$, $\mathbf{h}$ and $\mathbf{r}$, or equivalently minimizing (8), and using the gradient descent proposed in [28], we obtain the following updating rules of the unconstrained

NMF (UNMF):

$$\mathbf{h}^T \leftarrow \frac{1}{F}\mathbf{1}_F^T \frac{\mathbf{S}}{\mathbf{R} \bullet \prod_p \widehat{\mathbf{S}}^p} \tag{28}$$

$$\mathbf{r} \leftarrow \frac{1}{N} \frac{\mathbf{S}}{\mathbf{H} \bullet \prod_p \widehat{\mathbf{S}}^p} \mathbf{1}_N \tag{29}$$

$$\mathbf{G}^p \leftarrow \mathbf{G}^p \frac{\boldsymbol{\nabla}_-^p}{\boldsymbol{\nabla}_+^p} \tag{30}$$

where before each update, we compute:

$$\widehat{\mathbf{S}} = \mathbf{H} \bullet \mathbf{R} \bullet \prod_p \widehat{\mathbf{S}}^p \text{ with } \widehat{\mathbf{S}}^p = \mathbf{W}^p \mathbf{G}^p \ ,$$

$$\boldsymbol{\nabla}_-^p = \mathbf{W}^{pT} \frac{\mathbf{S}}{\widehat{\mathbf{S}} \bullet (\mathbf{W}^p \mathbf{G}^p)} \text{ and } \boldsymbol{\nabla}_+^p = \mathbf{W}^{pT} \frac{1}{\mathbf{W}^p \mathbf{G}^p}$$

In all the above equations, the operator $\div$ is element-wise. After each corresponding update, $\mathbf{G}$ and $\mathbf{r}$ should be normalized, such that most of the energy information is found in $\mathbf{h}$. As for VA-SFFHMM, $\mathbf{r}$ needs to be smooth, and was further modeled as $\mathbf{r} = \mathbf{W}^R \mathbf{G}^R$, where $\mathbf{W}^R$ is a dictionary of smooth frequency elements, in a similar fashion as the filter parts in [21]. However, for clarity, the corresponding updates are not detailed in this article.

B. Temporal constraints and sequence estimation through sparsity

In comparison with the VA algorithm, the UNMF provides estimates of $\mathbf{G}$ which are full matrices, while the FHMM structure requires that $\mathbf{G}$ is sparse, with temporal constraints shaping its structure: $\mathbf{g}_n^p$ should contain only one non-zero value, and the evolution from $\mathbf{g}_n$ to $\mathbf{g}_{n+1}$, $\forall n$ should be governed by the HMM transition probabilities.

In order to impose both the sparsity and the temporal constraint, an annealing strategy is proposed [21]: after each update (30), the estimates of $\mathbf{G}^p$ are re-weighted by a function that concentrates the energy around a single value, reducing the spread over the iteration of the estimation algorithm.

Sparsity: $\forall p, n$, let $\boldsymbol{\omega}_n^p$ be a vector of weights such that:

$$\omega_{kn}^p = \exp -\frac{(k - \mu_n^p)^2}{2\sigma_i^2} \tag{31}$$

where μ_n^p is the index number where the energy is focused and σ_i is the allowed spread, which depends on the iteration number i. When $\sigma \rightarrow 0$, then the weights are all 0, except for $k = \mu_n^p$. Using such

weights, with decreasing values of σ_i during the estimation of $\mathbf{g}_n^p$ allows to concentrate the estimated energies towards a single "state", by re-weighting the UNMF estimates as follows:

$$\mathbf{g}_n^p \leftarrow \boldsymbol{\omega}_n^p \bullet \mathbf{g}_n^p \tag{32}$$

μ_n^p is not known in advance, and the barycenter of the indices k, weighted by their values g_{kn}^p, can be used as an estimate for the current value of μ_n^p, at any given iteration of the algorithm[4]. Furthermore, in order to limit octave errors in the source part ($p = 0$), the weights for the barycenter calculus also bear a component to penalize higher F0s:

$$\mu_n^0 \triangleq \sum_k \frac{g_{kn}^0 (K-k)^2}{\sum_l g_{ln}^0 (K-l)^2} k \text{ and } \mu_n^p \triangleq \sum_k \frac{g_{kn}^p}{\sum_l g_{ln}^p} k \tag{33}$$

Smoothness: The above discussion allows to constrain the estimation of $\mathbf{G}$ to get closer to the desired sparsity. Furthermore, it provides a convenient framework to impose some smoothness in the state transition. In order to reproduce a behavior similar to the HMM model for each formant frequency track, we propose to limit the variations of the sequences $\{\mu_n^p\}$ by applying a median filter of order $Q = 10$ to each of them:

$$\widetilde{\mu}_n^p \leftarrow \text{median}\{\mu_{n-Q}^p, \ldots, \mu_{n+Q}^p\} \tag{34}$$

Such a smoothing scheme closely corresponds to the FHMM structure discussed in Sect. II-B. There are however several differences: the chosen smoothing scheme also assumes frequency proximity, such that the elements in each formant and F0 basis need to be ordered by frequency value f_{kp}. In addition, in the formant dictionaries, the elements are then ordered by amplitude ρ_{kp}. The proposed smoothing technique therefore also implies, to a certain extent, a stronger possibility to stay on a given formant frequency and change the bandwidth, rather than changing the frequency.

The median filtering also imposes constraints that can be considered as extending the temporal condition farther than the Markov conditioning, the latter being limited to consecutive frames. The sequences therefore tend to be smoother with the Source/Filter Sparse NMF (SFSNMF) approximation presented in this section, than with VA-SFFHMM. The proposed SFSNMF is close to works on sparsity such as the iterative reweighted ℓ_1-norm (IRWL1) minimization [29] or the smoothed $\mathcal{L}^0$ norm [30]. In [31], the author also proposes constraints in the cost function aiming at promoting the temporal smoothness and sparsity. It is however different from the proposed algorithms because we desire to obtain smooth

[4] It is worth noting that if $\mathbf{g}_n^p$ was the histogram of values drawn from a normal distribution, with parameters μ and σ, then the proposed barycenter would be an estimate of the mean μ.

frequency lines, and not smooth amplitude evolution. Furthermore, the sparseness proposed in [31] was found, in our early experiments, difficult to set up, and not well suited in our case because we want to impose that, within each formant dictionary, only one coefficient is non-null.

The final SFSNMF algorithm is given in Alg. 2. We set the number of iterations to $I = 80$. The complexity of this algorithm, per iteration, is smaller than that of the VA-SFFHMM, of the order of $\mathcal{O}(NPKF)$, mainly because there is no need for a forward-backward procedure. However, the annealing process requires more iterations so as to lead to reliable results. At last, the desired sequence of states (13) is here approximated, after the termination of the iterative algorithm, by:

$$\widehat{\mathbf{k}}^p \approx \widetilde{\mu}_n^p \tag{35}$$

Algorithm 2 SFSNMF: Iterative NMF algorithm to estimate the formant tracks

Initialize $\mathbf{G}^p$, $\forall p$, with positive random values
for i from 0 to I **do**
  for p from 0 to P **do**
    Update $\mathbf{G}^p$ using (30)
    Compute $\{\mu_n^p\}_n$ Eq. (33)
    Smooth the sequence with Eq. (34)
    Reweight $\mathbf{G}^p$, Eq. (32), and normalize.
    Update recording conditions $\mathbf{r}$ (29)
    Update energy compensation $\mathbf{h}$ (28)
  end for
end for
for p from 0 to P **do**
  $\widehat{\mathbf{k}}^p \leftarrow \widetilde{\mu}_n^p$
end for

V. Experiments

In this section, the databases are first described. The model meta-parameters are then discussed. At last the formant track estimations as well as the re-synthesis possibilities are commented.

A. Databases

The proposed algorithm was first evaluated on the Hillenbrand vowel database [2] (`Hill`), consisting of 1668 `/hVd/` utterances, where `V` is a vowel from the "phonetic" set {`ae`, `ah`, `aw`, `eh`, `er`, `ei`, `ih`, `iy`, `oa`, `oo`, `uh`, `uw`}. The sampling rate is 16 kHz. For each utterance, the first 3 formants are annotated on 8 "reliable" locations.

Another annotated database (`Deng`) was proposed in [32]. It consists of about 500 speech excerpts from the TIMIT database [33], for which the first four formants (or VTR) of each frame have been annotated. The annotation method was somewhat different from the `Hill` corpus, since each frame (with time-frequency analysis step $\delta = 10$ms or 160 samples) is annotated, with hand-corrected verifications. For some phoneme categories, especially unvoiced ones, the annotated formants are difficult to interpret, as they correspond to no spectral energy peak. However, it is interesting to compare our method to more classical methods on that database, since it provides excerpts which are longer and cover more phonetic content.

For the proposed algorithm, in particular for successful F0 estimations, a good frequency resolution is required, and the STFTs are therefore computed on 64ms-long windows (1024 samples), weighted by a Hann window, every $\delta = 16$ms (256 samples) for `Hill` and $\delta = 10$ms for `Deng`. We compare each "ground-truth" (GT) formant track $\{f^0_{qn}\}_n$ to each estimated formant track $\{f_{pn}\}_n$. The tracks are compared on the frames for which the annotation is available (8 for `Hill`, all the frames for `Deng`). For each file v, the standard deviation between the different estimated tracks $\{f^v_{pn}\}_n$ and GT tracks $\{f^{0,v}_{qn}\}_n$ are computed as the mean squared error, with frequencies expressed in Mel: $\epsilon^2_{qpv} = \frac{1}{N}\sum_n |f^v_{pn} - f^{0,v}_{qn}|^2$. Additionally, we also computed the mean absolute frequency errors as provided in [32], for comparison purposes[5].

B. Model meta-parameter estimation

As seen in Sect. II-C, many parameters such as the $\mathbf{W}^{F_p}$ are already set by fixed values, derived from the source/filter production model. However, some parameters remain to be estimated: the prior and transition probabilities for the states of each chain on the SFFHMM are separately learned on the annotated database, rather than re-estimated on each file, as explained below.

[5] It is worth noting that since the GT formant tracks are based on LPC estimates of the pole frequencies, they can be directly compared to our estimates, even if they may not exactly correspond to spectral peaks of the spectral envelope.

In order to give equal importance to all the possible F0 and formant frequencies, the prior distribution for each HMM chain states is uniform: $p(g^p_{k1}) = \pi^p_k = 1/K$, $\forall k$. These priors could also be learned to better fit a specific speaker.

As for the transition probabilities, $p(g^p_{kn} = j | g^p_{k(n-1)} = i) = P^p_{ij}$, a Laplace distribution is assumed such that P^p_{ij} only depends on the difference of the logarithm in base 2 between the formant or F0 frequency of each state, owing to the assumption that human auditory perception is based on a logarithmic frequency scale:

$$P^p_{ij} \propto \exp - \frac{|\log_2 f^p_j - \log_2 f^p_i|}{\lambda_p} \tag{36}$$

where λ_p depends on whether (36) models F0 or formant transitions, as well as on δ. Let δ_0 be the analysis step of the annotated formant frequencies. We first learn the parameter λ_{p,δ_0} from the annotations of `Hill`, with, on average, a step of $\delta_0 = 26$ms between the annotated formants, and we obtain $\lambda_{p,\delta_0} \approx 0.031$. Let δ_1 be another analysis step (for instance to analyze `Hill`, we use $\delta_1 = 16$ms). We then approximate the desired parameter by adapting the learned parameter such that $\lambda_{p,\delta_1} \approx \frac{\delta_0}{\delta_1} \lambda_{p,\delta_0} = 0.05$. Likewise, for $p = 0$, we have $\lambda_{0,\delta_0} \approx 0.278$ and $\lambda_{0,\delta_1} \approx 0.45$.

Furthermore, different numbers of inner loop iterations for the EM algorithm, *i.e.* the number of updates between the HMM chains, were tested. Most of the parameters are fixed, notably the dictionaries, and only the energy and recording condition filter need to be estimated. A limited number of iterations is needed, and 3 such iterations already provide very satisfying results. The reported results correspond to 3 iterations over the updates of all the parameters, for all the chains. Stopping conditions on the convergence of $\langle \mathbf{G} \rangle$ can also be used [23].

The annealing parameter σ_i in the SFSNMF model is chosen such that it follows an exponential decrease over the iterations, from $\sigma_0 = K^2$ to $\sigma_I = 3^2$, where I is the number of iterations:

$$\sigma_i = \exp\left(\log \sigma_0 + (\log(\sigma_I) - \log(\sigma_0)) \frac{i}{I}\right) \tag{37}$$

During the first iterations, the estimation is therefore almost unconstrained, with weights close to 1 everywhere. The last value σ_I corresponds to allowing significant energy for only 3 atoms away from the estimated state, at the end of the estimation process. At the end of the iterations, for each frame and each formant, there is therefore energy only around one component, as desired.

C. Formant track estimation

We ran the VA-SFFHMM and the SFSNMF algorithms, respectively Alg. 1 and Alg. 2, on the aforementioned databases. Fig. 4 and 5 show the results obtained for the databases `Hill` and `Deng`,

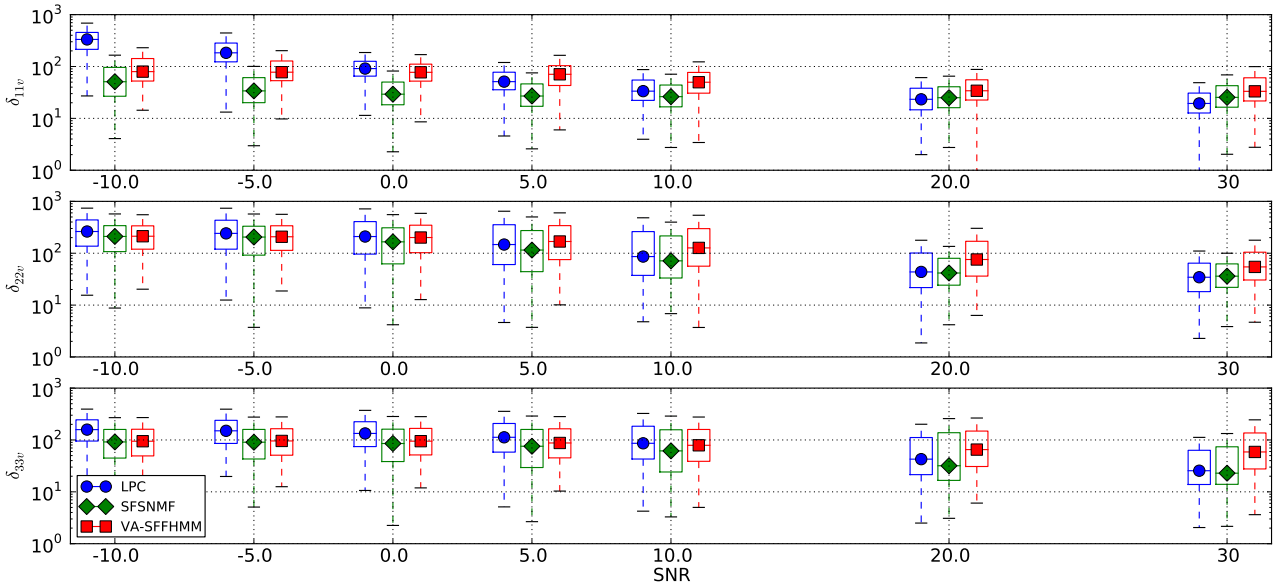


Fig. 4. Box-plots of the estimation errors ϵ_{ppv} for each of the 3 GT formant tracks on `Hill`, for the LPC, SFSNMF and SFFHMM formant tracks and with various SNR conditions. The markers represent the medians, the rectangles the second and third quartiles, the whiskers show the extent of the errors, and the circles are outliers.

respectively. For the SFSNMF, on `Hill`, as in [21], formant dictionaries $\mathbf{W}^p$ were generated for $p = 1$ to 5, with $G_F = 60$ and $G_R = 5$, plus the neutral vector, resulting in a total of $K = 301$ vectors in each matrix $\mathbf{W}^p$. However, on `Deng`, and for the VA-SFFHMM algorithm on both datasets, allowing more formant chains, from $p = 1$ to 7, was preferred, in particular because VA-SFFHMM is less flexible than SFSNMF, and therefore needs more degrees of flexibilities included directly in the model. In that case, G_F and G_R were respectively set to 40 and 10, with a total of $K = 401$ vectors per dictionary. In this section, we report the results for these values.

The proposed algorithms were compared to other systems: on `Hill`, we used a simple `LPC` system, which estimates the formant frequencies from the AR(14) polynomial roots, by sorting these frequencies by increasing frequency values. On `Deng`, the results provided in [32] for two systems are re-used: first, the formant estimation of `WaveSurfer` [34] involves an LPC estimation, followed by a dynamic programming step that forms the formant tracks with a penalization that takes into account both the frequency and amplitude of the LPC roots [5], and second, the `MSR` also involves LPC coefficients, but with a Kalman filter approach to estimate the formant tracks [12]. These approaches benefit from very efficient algorithms, and their computational times are faster or close to real time.

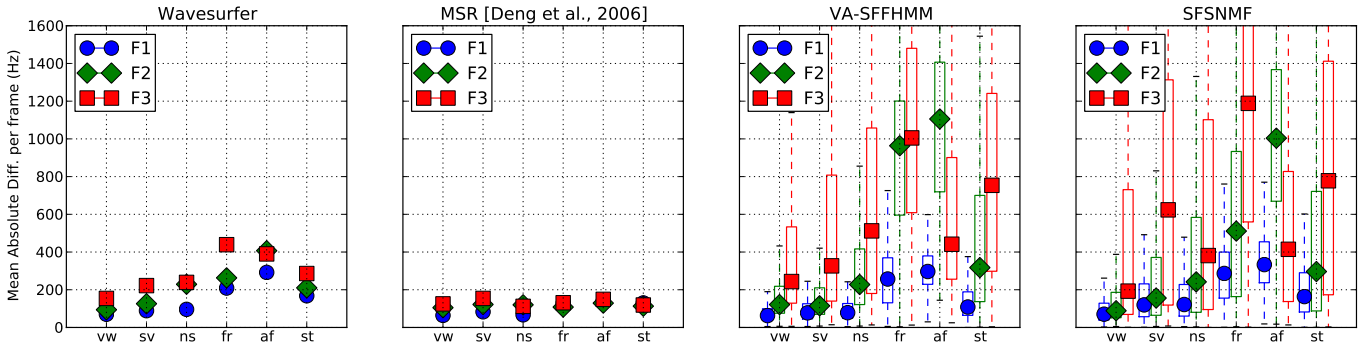


Fig. 5. Mean absolute estimation error, for 3 GT formant tracks on `Deng`. The wavesurfer and MSR's results are from [32], to be compared to our VA-SFFHMM and SFSNMF results, shown as box-plots, as in Fig. 4. The results are shown per categories: vowels (vw), semivowels (sv), nasal (ns), fricatives (fr), affricatives (af), stops (st).

On `Hill`, the proposed algorithms were tested for various Signal-to-Noise-Ratios (SNR), by adding an additive white Gaussian noise at different levels. According to Fig. 4, one can see that the noise mostly affects the LPC algorithm, as well as VA-SFFHMM, for which the performance gradually increases with the SNR, especially for F1 and F2. On the contrary, the effect is negligible for SFSNMF, thanks to the flexibility of the model. For F3, the variations of the estimation errors are however such that the differences between the different SNR conditions are not significant, except for the LPC estimation.

Furthermore, it is worth noting that an algorithm simply returning, for each formant, the mean formant value over the database would obtain average errors ϵ_{ppv} of around 100, 200 and 90, respectively for the estimation of F1, F2 and F3 ($p = 1, 2, 3$). In Fig. 4 it is therefore shown that the proposed algorithms SFSNMF and VA-SFFHMM both return very good results at high SNRs for F1 and F2. For F1, SFSNMF also performs very well with respect to this baseline. However, for F3, this simple method performs as well as the other methods. The estimation of F3 indeed requires more care and further investigation, as can be seen on the example from `Deng`, in Fig. 6.

On `Deng`, the proposed algorithms also provide good results for several phoneme categories: the vowels, semivowels, nasals and stops (Fig. 5). However, like the Wavesurfer estimations [32], ours seem to fail on the fricatives and affricatives. The choice of the annotator for these categories in particular would require a specific treatment. Indeed, the annotated first VTR does not always correspond to a spectral peak. Such formant tracks were mainly interpolated between two "clear" (voiced) formant track sections. The proposed models include, in theory, formant contributions that do not lead to peaks in the spectrum (the flat spectra in each of the $\mathbf{W}^p$), but in our experiments, these states in the formant

sequences were hardly chosen by the algorithms to explain the data.

As concerns the F0 estimation, an additional issue encountered in our model is the difficulty to fix values for the transition probabilities, so as to enable the unvoiced component in the source part to be chosen. With the SFFHMM, instead of the unvoiced component, low F0s or high F0s are often taken. Furthermore, informally, the model with the noise threshold introduced in Sect. III-B was found to provide better fitting decompositions than those of the model without that threshold. From inspecting the power spectra, it also turns out unvoiced frames do not completely follow the assumed model of a white noise convolved with the vocal tract filter: the spectra are indeed not as smooth as expected. Further investigation, with some pre-processing steps, should tell how to address this issue.

The average computation times are about 30 times real time for the VA-SFFHMM algorithm, and around 50 times for the SFSNMF one. While the LPC estimation is faster and with usually good results, there are several other advantages that make the proposed framework attractive: first, the signal model allows to take into account the complete observed signal, and not just features. The observation likelihood indeed involves here the original signal or an invertible transform, while most of the other approaches rely on extracted features, such as the LPC coefficients, which are only estimations to some aspects of the signal (for instance, with LPC coefficients, the spectral envelope). The proposed "analysis-by-synthesis" scheme therefore provides a natural way of re-synthesizing the speech signal, as shown below. Furthermore, in particular for the SFSNMF, the model is easily extendable to the case where there are several concurrent speakers at once. Such an extension for the VA-SFFHMM would however require more effort, using for instance a different observation likelihood, or a mixture-maximization model [18].

D. Speech re-synthesis

At last, another way of assessing the performance in terms of representation of the proposed model is to re-synthesize the speech utterances. Indeed, contrary to LPC parameters, the model parameters completely characterize the speech signal and allow for synthesis without use of the original signal. In order to resynthesize the signal, the model parameters are first estimated. The source signal is then synthesized using the estimated F0 track, and subsequently filtered in the frequency domain by the estimated filter part, frame by frame. The synthesized signal is at last obtained by an overlap-and-add procedure.

Informally, the speech signals resulting from the VA-SFFHMM parameters $\{\mathbf{r}, \mathbf{h}, \mathbf{K}\}$ are clear and the content is understandable: as shown on Fig. 6, the spectrogram reconstructed from the estimated F0 and formant tracks by the VA-SFFHMM algorithm is very close to the original one. In comparison, from the

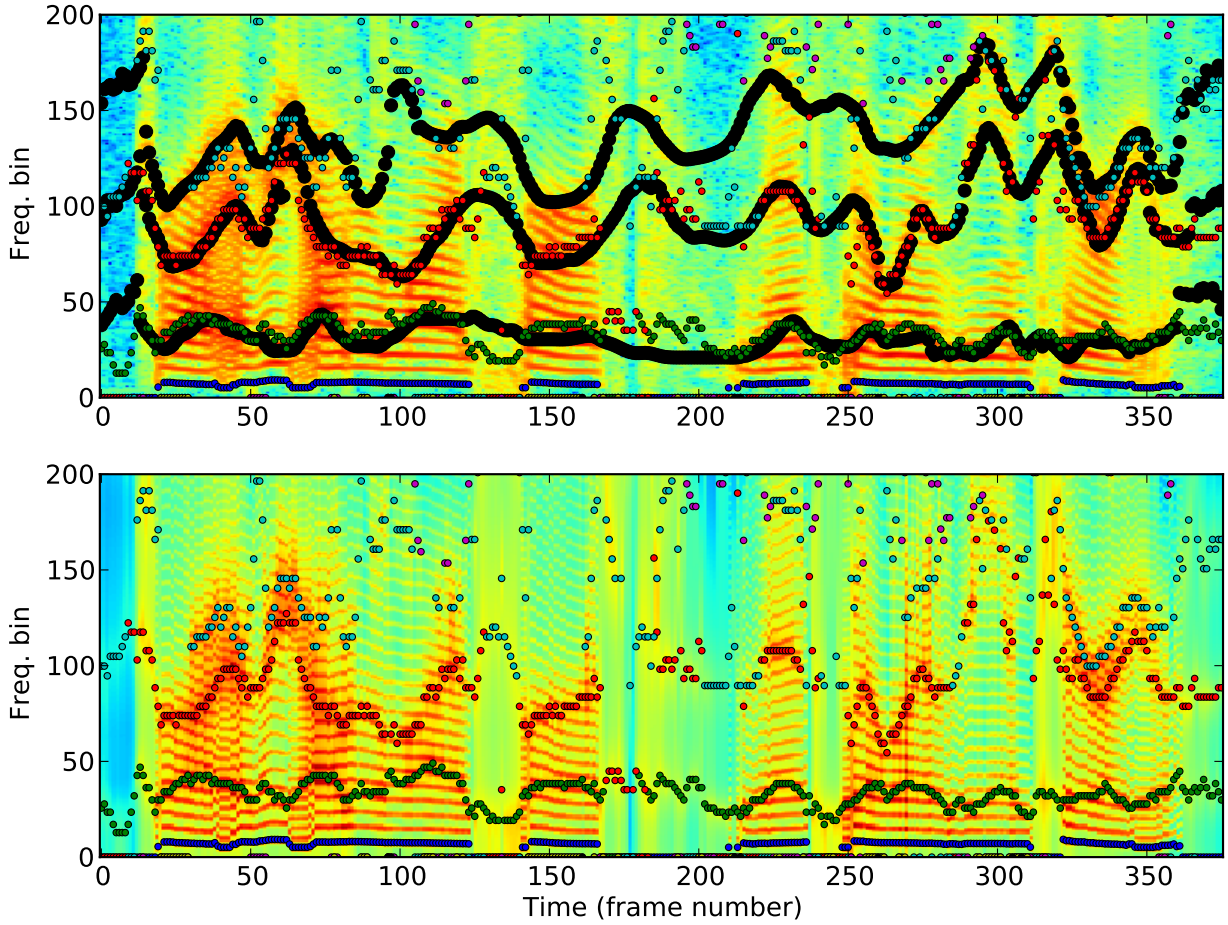


Fig. 6. Top pane: original spectrogram, ground truth formant tracks (black dots) with estimated F0 and formant superimposed (colored dots). Bottom pane: VA-SFFHMM reconstructed spectrogram and track estimates for a speech utterance from TIMIT (`train/dr1/mkls0/sx267`: "Draw every outer line first, then fill in the interior.").

SFSNMF parameters, the general intonation is usually kept, but with a phonetic content which is more difficult to identify. It is actually more difficult to re-synthesize from these latter parameters, because the global energy and the presence of flat spectra in the dictionaries requires to store the energies of these elements $\{g^p_{Kn}\}_n$ in addition to the other parameters.

Since the estimated parameters completely characterize the signals, it is also very easy to modify the F0, formants or energies in order to modify the speech signal. It also provides a way to learn models for a fully parametric speech synthesis system. Sound examples and software to reproduce some results are available at http://www.durrieu.ch/research/sffhmm2012.html.

VI. Conclusion

In order to model speech signals and in particular to track the underlying formant frequencies, a factorial HMM was combined with a source/filter speech production model. The desired parameters and frequency sequences are estimated through two proposed algorithms: due to the intractable nature of the original FHMM, a variational approximation approach is first derived. Another approximation, based on NMF methodology, is then investigated.

Both algorithms proved to be stable enough to provide reliable estimations, as shown by our experiments on publicly available databases. The results are comparable with other state-of-the-art vocal tract resonance tracking. The proposed model also provides new insights in speech models, relying on a strong theoretical framework.

Future works should aim at improving the algorithms, in particular with respect to tracking unvoiced segments. Parameters such as the formant frequency ranges, the number of formant HMM chains and the transition probabilities between the different formant states will also be tested. Extensions of the proposed model include background noise and multiple concurrent speaker modeling, thus improving the robustness of the systems under various challenging situations. Further reducing the computational cost of the algorithms should also be sought after: the VA-SFFHMM algorithm could for instance be multi-threaded such that all the formant and F0 tracks are processed in parallel, at the cost, however, of breaking the optimization procedure. At last, the extracted parameters could also be used in other applications, such as speech synthesis, speech coding, or speech recognition, hence providing features that depend more on the production process, and less on speaker or channel conditions.

Appendix A
VA-SFFHMM: M-step derivations

In this appendix, the updating rules (25) and (26) are derived. The energy h_n and the recording filter $\mathbf{r}$ need to be re-estimated each time a new variational parameter set is updated, since all these parameters are inter-dependent. This M-step can also be used to classically update quantities such as $\mathbf{C}$ or the transition probabilities $\mathbf{P}$, but we did not consider these updates in this work. The interested reader can refer to [23] for more details on updating these parameters.

The EM criterion (24) can be written as:

$$\begin{aligned}\mathcal{Q}(\theta|\theta^{\text{old}}) = &-\frac{N}{2}\log((2\pi)^F|\mathbf{C}|)\\ &-\frac{1}{2}\sum_n(\tilde{\mathbf{s}}_n-\tilde{h}_n-\tilde{\mathbf{r}})^T\mathbf{C}^{-1}(\tilde{\mathbf{s}}_n-\tilde{h}_n-\tilde{\mathbf{r}})\\ &+\sum_n(\tilde{\mathbf{s}}_n-\tilde{h}_n\mathbf{1}_F-\tilde{\mathbf{r}})^T\mathbf{C}^{-1}\left(\sum_p\widetilde{\mathbf{W}}^p\langle\mathbf{g}_n^p\rangle\right)\\ &-\frac{1}{2}\sum_{p,q}\text{tr}\left(\widetilde{\mathbf{W}}^{pT}\mathbf{C}^{-1}\widetilde{\mathbf{W}}^q\langle\mathbf{g}_n^q\mathbf{g}_n^{pT}\rangle\right)+\texttt{CST}\end{aligned} \tag{38}$$

where CST is a term independent from the desired parameters. Note that CST actually depends on $\mathbf{P}^p$ and π^p, and can be developed in case these parameters are also to be estimated. As in [23], to estimate the desired parameters, in our case $\mathbf{h}$ and $\mathbf{r}$, the derivatives against these vectors need to be computed:

$$\frac{\partial\mathcal{Q}}{\partial\tilde{h}_n} = (\tilde{\mathbf{s}}_n-\tilde{\mathbf{r}}-\tilde{h}_n\mathbf{1}_F-\sum_p\widetilde{\mathbf{W}}^p\langle\mathbf{g}_n^p\rangle)^T\mathbf{C}^{-1}\mathbf{1}_F \tag{39}$$

$$\frac{\partial\mathcal{Q}}{\partial\tilde{\mathbf{r}}} = \sum_n\left[(\tilde{\mathbf{s}}_n-\tilde{\mathbf{r}}-\tilde{h}_n\mathbf{1}_F)\mathbf{C}^{-1}-(\sum_p\widetilde{\mathbf{W}}^p\langle\mathbf{g}_n^p\rangle)^T\mathbf{C}^{-1}\right] \tag{40}$$

Setting these partial derivatives to 0 respectively leads to the update formulas (25) and (26).

APPENDIX B
VA-SFFHMM: E-STEP DERIVATIONS

The E-step mainly provides the posterior probabilities $\langle\mathbf{G}\rangle$ through a classical forward-backward procedure [22] for each chain p, replacing the likelihood with $\{\tilde{\mathbf{t}}_n^p\}_n$. The estimation of these variational parameters is therefore required, and the derivations are detailed in this appendix.

The KL divergence (18) between the true pdf p and the structured variational approximation pdf q is similar to that of [23]:

$$\begin{aligned}D_{\text{KL}}(q|p) = &-\log Z_q+\log Z+\sum_{n,p}\langle\mathbf{g}_n^p\rangle^T\tilde{\mathbf{t}}_n^p\\ &+\frac{1}{2}\sum_n(\tilde{\mathbf{s}}_n-\tilde{h}_n-\tilde{\mathbf{r}})^T\mathbf{C}^{-1}(\tilde{\mathbf{s}}_n-\tilde{h}_n-\tilde{\mathbf{r}})\\ &-\sum_n(\tilde{\mathbf{s}}_n-\tilde{h}_n-\tilde{\mathbf{r}})^T\mathbf{C}^{-1}\left(\sum_p\widetilde{\mathbf{W}}^p\langle\mathbf{g}_n^p\rangle\right)\\ &+\frac{1}{2}\sum_{n,p}\text{tr}\left(\widetilde{\mathbf{W}}^{pT}\mathbf{C}^{-1}\widetilde{\mathbf{W}}^p\text{diag}\left(\langle\mathbf{g}_n^p\rangle\right)\right)\\ &+\frac{1}{2}\sum_{n,p,q\neq p}\text{tr}\left(\widetilde{\mathbf{W}}^{pT}\mathbf{C}^{-1}\widetilde{\mathbf{W}}^q\langle\mathbf{g}_n^q\mathbf{g}_n^{pT}\rangle\right)\end{aligned} \tag{41}$$

where Z is the normalizing constant for pdf p, which is independent from the variational parameters.

The partial derivatives with respect to the variational parameters $\tilde{\mathbf{t}}_n^p$ are given by:

$$\frac{\partial D_{\mathrm{KL}}}{\partial \tilde{\mathbf{t}}_n^p} = \sum_{n,p} \left[\frac{\partial \langle \mathbf{g}_n^p \rangle}{\partial \tilde{\mathbf{t}}_n^p} \right]^T \Big[\tilde{\mathbf{t}}_n^p - \widetilde{\mathbf{W}}^{pT} \mathbf{C}^{-1} (\tilde{\mathbf{s}}_n - \tilde{h}_n - \tilde{\mathbf{r}}) + \frac{1}{2} \Delta^p + \sum_{q \neq p} \widetilde{\mathbf{W}}^{pT} \mathbf{C}^{-1} \widetilde{\mathbf{W}}^q \langle \mathbf{g}_n^q \rangle \Big] \tag{42}$$

where Δ^p was described in Sect. III-C. Setting, for all n and p, the elements in the brackets to 0 leads to the Eq. (23).

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions which helped to improve this document.

REFERENCES

[1] G. Fant, *Acoustic Theory of Speech Production*. Mouton, 1970.

[2] J. Hillenbrand, L. Getty, M. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *JASA*, vol. 97, pp. 3099–3111, 1995.

[3] L. Deng, D. Yu, and A. Acero, "Structured speech modeling," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1492 – 1504, September 2006.

[4] Y. Iribe, S. Manosavan, K. Katsurada, R. Hayashi, C. Zhu, and T. Nitta, "Improvement of animated articulatory gesture extracted from speech for pronunciation training," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 25-30 2012.

[5] D. Talkin, "Speech formant trajectory estimation using dynamic programming with modulated transition costs," *Journal of the Acoustical Society of America*, vol. 82, no. S1, pp. S55–S55, 1987. [Online]. Available: http://link.aip.org/link/?JAS/82/S55/1

[6] L. Rabiner and R. Schafer, *Digital processing of speech signals*, ser. Prentice-Hall signal processing series. Prentice-Hall, 1978.

[7] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot Int.*, vol. 5, no. 9/10, pp. 341–345, 2001.

[8] P. Zolfaghari and T. Robinson, "Formant analysis using mixtures of gaussians," in *Proc. of the International Conference on Spoken Language*, vol. 2, oct 1996, pp. 1229 –1232 vol.2.

[9] E. Ozkan, I. Ozbek, and M. Demirekler, "Dynamic speech spectrum representation and tracking variable number of vocal tract resonance frequencies with time-varying dirichlet process mixture models," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 8, pp. 1518 –1532, nov. 2009.

[10] P. Depalle, G. Garcia, and X. Rodet, "Tracking of partials for additive sound synthesis using hidden markov models," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Minneapolis, Minnesota, USA, 27-30 April 1993, pp. 225–228.

[11] I. Ozbek and M. Demirekler, "Tracking of vocal tract resonances based on dynamic programming and kalman filtering," in *Proc. of the IEEE Signal Processing, Communication and Applications Conference*, april 2008, pp. 1 –4.

[12] L. Deng, I. Bazzi, and A. Acero, "Tracking vocal tract resonances using an analytical nonlinear predictor and a target-guided temporal constraint," in *proc. of Eurospeech*, Geneva, Switzerland, 2003.

[13] L. Deng, L. J. Lee, H. Attias, and A. Acero, "A structured speech modelwith continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances," in *proc. of International Conference on Acoustics, Speech and Signal Processing*, May 17-21 2004.

[14] J. Vargas and S. McLaughlin, "Cascade prediction filters with adaptive zeros to track the time-varying resonances of the vocal tract," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 1 –7, jan. 2008.

[15] G. Richard, M. Goirand, D. Sinder, and J. Flanagan, "Simulation and visualization of articulatory trajectories estimated from speech signals," in *ASVA*, 1997.

[16] N. Moal and J. Fuchs, "Estimation de l'ordre et identification des paramètres d'un processus ARMA," in *16ème Colloque sur le traitement du signal et des images*. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images, 1997.

[17] S. Roweis, "One microphone source separation," *Advances in Neural Information Processing Systems*, vol. 13, pp. 793–799, 2001.

[18] M. Wohlmayr, M. Stark, and F. Pernkopf, "A probabilistic interaction model for multipitch tracking with factorial hidden Markov models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 799 –810, may 2011.

[19] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden markov model for polyphonic audio representation and source separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, NY, USA, Oct. 18-21 2009.

[20] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, March 2009.

[21] J.-L. Durrieu and J.-P. Thiran, "Sparse non-negative decomposition of speech power spectra for formant tracking," in *proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 22-27 2011.

[22] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.*, vol. 41, no. 1, p. 164171, 1970.

[23] Z. Ghahramani and M. Jordan, "Factorial hidden Markov models," *Machine Learning*, vol. 29, no. 2, pp. 245–273, 1997.

[24] D. Klatt and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *JASA*, vol. 87, pp. 820–857, 1990.

[25] R. Schafer and L. Rabiner, "System for automatic formant analysis of voiced speech," *JASA*, vol. 47, p. 634, 1970.

[26] L. Benaroya, "Séparation de plusieurs sources sonores avec un seul microphone," Ph.D. dissertation, Université de Rennes 1, 2003.

[27] B. Rivet, L. Girin, and C. Jutten, "Log-Rayleigh distribution: A simple and efficientstatistical representation of log-spectral coefficients," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 796 – 802, March 2007.

[28] D. Lee and H. Seung, "Algorithms for Non-negative Matrix Factorization," *Advances in Neural Information Processing Systems*, pp. 556–562, 2001.

[29] M. Duarte, M. B. Wakin, and R. G. Baraniuk, "Wavelet-domain compressive signal reconstruction using a hidden Markov tree model," in *proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, Nevada, USA, March 31-April 4 2008.

[30] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed $\mathcal{L}^0$ norm," *IEEE TSP*, vol. 57, no. 1, pp. 289–301, 2009.

[31] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[32] L. Deng, X. Cui, R. Pruvenok, Y. Chen, S. Momen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, 14-19 May 2006.

[33] J. S. Garofolo and et al., "TIMIT acoustic-phonetic continuous speech corpus," Linguistic Data Consortium, Philadelphia, 1993.

[34] K. Sjölander and J. Beskow, "WaveSurfer - an open source speech tool," in *proc. of the 6th International Conference on Spoken Language Processing*, Beijing, China, 2000, pp. 464–467.

Jean-Louis Durrieu was born on August 14th, 1982, in Saint-Denis, Reunion Island, France. He received the State Engineering degree and the Ph.D. degree, in the field of audio signal processing, from TELECOM ParisTech (formerly ENST), Paris, France, in 2006 and in 2010, respectively. His main research interests are statistical models for audio signals, with applications to language learning technologies, musical audio source separation and music information retrieval. He is currently a Post-Doctoral Scientist in the Signal Processing Laboratory (LTS5) at the École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.

Jean-Philippe Thiran (S'91 - M'98 - SM'03) received the Elect. Eng. and Ph.D. degrees from the Université catholique de Louvain, Louvain-la-Neuve, Belgium, in 1993 and 1997, respectively. Since 2004 he has been a Professor at the Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, and director of the Signal Processing Laboratory (LTS5) at EPFL. Since 2011 he also holds a part time Associate Professor position with the University Hospital Center (CHUV) and University of Lausanne (UNIL). His current scientific interests include image segmentation, prior knowledge integration in image analysis, partial differential equations and variational methods in image analysis and multimodal signal processing, with many applications including remote sensing, medical imaging and human-computer interactions. Dr. Thiran was Co-editor-in-Chief of the "Signal Processing" international journal from 2001 to 2005. He is currently an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING. Among many other scientific duties, he was the General Chairman of the 2008 European Signal Processing Conference (EUSIPCO 2008), a tutorial co-chair of the IEEE Int. Conf. Image Processing in 2011 (ICIP-2011) and will be the technical co-chair of ICIP-2015. He is a Senior Member of the IEEE, member of the IVMSP technical committees of the IEEE Signal Processing Society.